Nick Bostrom
Cuando nuestras computadoras se vuelven más inteligentes que nosotros

Yo trabajo con un grupo de matemáticos, filósofos y científicos informáticos, y nos juntamos para pensar en el futuro de la inteligencia artificial, entre otras cosas. Algunos piensan que estas cosas son una especie de ciencia ficción alocadas y alejadas de la verdad. Pero bueno, me gusta sugerir que analicemos la condición humana moderna. Así es cómo tienen que ser las cosas.

Pero si lo pensamos, en realidad acabamos de llegar a este planeta, nosotros, la especie humana. Piensen que si la Tierra hubiera sido creada hace un año, entonces la raza humana solo tendría 10 minutos de edad y la era industrial habría empezado hace dos segundos. Otra forma de abordar esto es pensar en el **PIB** mundial en los últimos 10 000 años, y de hecho, me he tomado la molestia de representarlo en un gráfico para Uds. Se ve así. Tiene una forma curiosa para ser normal. Y seguro que no me gustaría sentarme en ella.

Preguntémonos ¿cuál es la causa de esta anomalía actual? Algunas personas dirán que es la tecnología. Ahora bien es cierto que la tecnología ha aumentado a lo largo de la historia, y en la actualidad avanza muy rápidamente --esa es la causa inmediata-- y por esto la razón de ser muy productivos hoy en día. Pero me gusta indagar más allá, buscar la causa de todo.

Miren a estos 2 caballeros muy distinguidos: Tenemos a Kanzi, que domina 200 unidades léxicas, una hazaña increíble, y a Ed Witten que desató la segunda revolución de las supercuerdas. Si miramos atentamente esto es lo que encontramos: básicamente la misma cosa. Uno es un poco más grande, y puede que también tenga algunos trucos más por la forma en que está diseñado, sin embargo, estas diferencias invisibles no pueden ser demasiado complicadas porque solo nos separan 250 000 generaciones de nuestro último ancestro común. Sabemos que los mecanismos complicados tardan mucho tiempo en evolucionar así que una serie de pequeños cambios nos lleva de Kanzi a Witten, de las ramas de árboles rotas a los misiles balísticos intercontinentales.

Así que parece bastante claro que todo lo que hemos logrado, y todo lo que nos importa, depende fundamentalmente de algunos cambios relativamente menores sufridos por la mente humana. Y el corolario es que, por supuesto, cualquier cambio ulterior que cambiara significativamente el fundamento del pensamiento podría potencialmente acarrear enormes consecuencias.

Algunos de mis colegas piensan que estamos muy cerca de algo que podría causar un cambio significativo en ese fundamento y que eso es la máquina superinteligente. La inteligencia artificial solía ser la integración de comandos en una caja, con programadores humanos que elaboraban conocimiento minuciosamente a mano. Se construían estos sistemas especializados y eran bastante útiles para ciertos propósitos, pero eran muy frágiles y no se podían ampliar. Básicamente, se conseguía solamente lo que se invertía en ellos. Pero desde entonces, hubo un cambio de paradigma en el campo de la inteligencia artificial.

Hoy, la acción gira en torno al aprendizaje máquina. Así que en lugar de producir características y representar el conocimiento de manera artesanal, creamos algoritmos que aprenden a menudo

a partir de datos de percepción en bruto. Básicamente lo mismo que hace el bebé humano. El resultado es inteligencia artificial que no se limita a un solo campo; el mismo sistema puede aprender a traducir entre cualquier par de idiomas o aprender a jugar a cualquier juego de ordenador en la consola Atari. Ahora, por supuesto, la **I.A. (Inteligencia Artificial)** está todavía muy lejos de tener el mismo poder y alcance interdisciplinario para aprender y planificar como lo hacen los humanos. La corteza cerebral aún esconde algunos trucos algorítmicos que todavía no sabemos cómo simular en las máquinas.

Así que la pregunta es, ¿cuánto nos falta para poder implementar esos trucos? Hace un par de años hicimos una encuesta entre los expertos de **I.A.** más importantes del mundo para ver lo que piensan, y una de las preguntas que hicimos fue, "¿En qué año crees que habrá un 50 % de probabilidad en elevar la inteligencia artificial al mismo nivel que la inteligencia humana?" Donde hemos definido ese nivel como la capacidad de realizar casi todas las tareas, al menos así como las desarrolla un humano adulto, por lo cual, un nivel real no solo dentro de un área limitada. Y la respuesta fue alrededor de 2040 o 2050, dependiendo del grupo de expertos consultados. Ahora, puede ocurrir mucho más tarde o más temprano, la verdad es que nadie lo sabe realmente.

Lo que sí sabemos es que el umbral en el procesamiento de información en una infraestructura artificial se encuentra mucho más allá de los límites del tejido biológico. Esto pertenece al campo de la física. Una neurona biológica manda impulsos quizá a 200 Hertz, 200 veces por segundo. mientras que incluso hoy, un transistor opera a la frecuencia de los gigahercios. Las neuronas propagan el impulso lentamente a lo largo de los axones, a máximo 100 metros por segundo. Pero en las computadoras, las señales pueden viajar a la velocidad de la luz. También hay limitaciones de tamaño, como el cerebro humano que tiene que encajar dentro del cráneo, pero una computadora puede ser del tamaño de un almacén o aún más grande. Así que el potencial de la máquina superinteligente permanece latente en la materia, al igual que el poder atómico a lo largo de toda la historia que esperó pacientemente hasta 1945. De cara a este siglo los científicos pueden aprender a despertar el poder de la inteligencia artificial y creo que podríamos ser testigos de una explosión de inteligencia.

Cuando la mayoría de la gente piensa en lo inteligente o lo tonto creo que tienen en mente una imagen más o menos así. En un extremo tenemos al tonto del pueblo, y lejos en el otro extremo, tenemos a Ed Witten o a Albert Einstein, o quien sea su gurú favorito. Pero creo que desde el punto de vista de la inteligencia artificial, lo más probable es que la imagen real sea la siguiente: Se empieza en este punto aquí, en ausencia de inteligencia y luego, después de muchos, muchos años de trabajo muy arduo, quizá finalmente lleguemos al nivel intelectual de un ratón, algo que puede navegar entornos desordenados igual que un ratón. Y luego, después de muchos, muchos más años de trabajo muy arduo, de mucha inversión, tal vez alcancemos el nivel de inteligencia de un chimpancé. Y luego, después de más años de trabajo muy, muy arduo alcancemos la inteligencia artificial del tonto del pueblo. Un poco más tarde, estaremos más allá de Ed Witten. El tren del progreso no se detiene en la estación de los Humanos. Es probable que más bien, pase volando.

Esto tiene profundas consecuencias, especialmente si se trata de poder. Por ejemplo, los chimpancés son fuertes. Un chimpancé es dos veces más fuerte y en mejor forma física que un hombre y, sin embargo, el destino de Kanzi y sus amigos depende mucho más de lo que hacemos los humanos que de lo que ellos mismos hacen. Una vez que hay superinteligencia, el destino de la humanidad dependerá de lo que haga la superinteligencia. Piensen en esto: la máquina inteligente es el último invento que la humanidad jamás tendrá que realizar. Las máquinas serán entonces mejores inventores que nosotros, y lo harán a escala de tiempo digital lo que significa básicamente que acelerarán la cercanía al futuro. Piensen en todas las tecnologías que tal vez, en su opinión, los humanos pueden desarrollar con el paso del tiempo: tratamientos para el envejecimiento, la colonización del espacio, nanobots autoreplicantes, mentes integradas en las computadoras, todo tipo de ciencia-ficción y sin embargo en consonancia con las leyes de la física. Todo esta superinteligencia podría desarrollarse y posiblemente con bastante rapidez.

Ahora, una superinteligencia con tanta madurez tecnológica sería extremadamente poderosa, y con la excepción de algunos casos sería capaz de conseguir lo que quiere. Nuestro futuro se determinaría por las preferencias de esta **I.A**. Y una buena pregunta es ¿cuáles son esas preferencias? Aquí se vuelve más complicado. Para avanzar con esto, debemos en primer lugar evitar el antropomorfismo. Y esto es irónico porque cada artículo de prensa sobre el futuro de la **I.A.** presenta una imagen como esta: Así que creo que tenemos que pensar de manera más abstracta, no según escenarios entretenidos de Hollywood.

Tenemos que pensar en la inteligencia como un proceso de optimización un proceso que dirige el futuro hacia un conjunto especifico de configuraciones. Un superinteligencia es un proceso de optimización realmente potente. Es muy bueno en el uso de recursos disponibles para lograr un estado óptimo y alcanzar su objetivo. Esto significa que no hay ningún vínculo necesario entre ser muy inteligente en este sentido, y tener una meta que para los humanos vale la pena o es significativa.

Por ejemplo, la **I.A.** podría tener el objetivo de hacer sonreír a los humanos. Cuando la **I.A.** está en desarrollo, realiza acciones entretenidas para hacer sonreír a su usuario. Cuando la **I.A.** se vuelve superinteligente, se da cuenta de que hay una manera más eficaz para lograr su objetivo: tomar el control del mundo e introducir electrodos en los músculos faciales de la gente para provocar sonrisas constantes y radiantes. Otro ejemplo, supongamos que le damos el objetivo de resolver un problema matemático difícil. Cuando la **I.A.** se vuelve superinteligente, se da cuenta de que la forma más eficaz para conseguir la solución a este problema es mediante la transformación del planeta en un computador gigante, para aumentar su capacidad de pensar. Y tengan en cuenta que esto da a la **I.A.** una razón instrumental para hacer cosas que nosotros no podemos aprobar. Los seres humanos se convierten en una amenaza, ya que podríamos evitar que el problema se resuelva.

Por supuesto, las cosas no tienen necesariamente que pasar de esa manera: son ejemplos de muestra. Pero lo importante, si crean un proceso de optimización muy potente, optimizado para lograr el objetivo X, más vale asegurarse de que la definición de X incluye todo lo que importa. Es una moraleja que también se enseña a través de varios mitos. El rey Midas deseaba convertir

en oro todo lo que tocaba. Toca a su hija y ella se convierte en oro. Toca su comida, se convierte en oro. Es un ejemplo relevante no solo de una metáfora de la codicia sino como ilustración de lo que sucede si crean un proceso de optimización potente pero le encomiendan objetivos incomprensibles o sin claridad.

Uno puede pensar: "Si una computadora empieza a poner electrodos en la cara de la gente bastaría simplemente con apagarla. En primer lugar, puede que no sea tan sencillo si somos dependientes del sistema por ejemplo: ¿dónde está el botón para apagar Internet? En segundo lugar, ¿por qué los chimpancés no tienen acceso al mismo interruptor de la humanidad, o los neandertales? Sin duda razones tendrían. Tenemos un interruptor de apagado, por ejemplo, aquí mismo. (Finge estrangulación) La razón es que somos un adversario inteligente; podemos anticipar amenazas y planificar en consecuencia, pero también podría hacerlo un agente superinteligente, y mucho mejor que nosotros. El tema es que no debemos confiar que podemos controlar esta situación.

Y podríamos tratar de hacer nuestro trabajo un poco más fácil digamos, poniendo a la **I.A.** en una caja, en un entorno de software seguro, una simulación de realidad virtual de la que no pueda escapar. Pero, ¿cómo podemos estar seguros de que la **I.A.** no encontrará un error? Dado que incluso los hackers humanos encuentran errores todo el tiempo, yo diría que probablemente, no podemos estar muy seguros. Así que desconectamos el cable ethernet para crear un espacio vacío, pero una vez más, al igual que los hackers humanos, podrían superar estos espacios usando la ingeniería social. Ahora mismo, mientras hablo, estoy seguro de que algún empleado, en algún lugar ha sido convencido para revelar los detalles de su cuenta por alguien que dice ser del departamento de IT.

Otros escenarios creativos también son posibles, por ejemplo si Ud. es la **I.A.**, puede hacer cambios en los electrodos de su circuito interno de seguridad para crear ondas de radio y usarlas para comunicarse. O tal vez fingir un mal funcionamiento, y cuando los programadores lo abren para entender qué está mal, al mirar el código fuente, ¡pum! ya empieza a manipular. O podría idear un programa tecnológico realmente ingenioso, y cuando lo implementamos, tener efectos secundarios ocultos planeados por la **I.A**. No debemos confiar en nuestra capacidad para mantener un genio superinteligente encerrado en su lámpara para siempre. Tarde o temprano, saldrá.

Creo que la solución es averiguar cómo crear una **I.A.** superinteligente para que incluso si, o cuando se escape, sea todavía segura para que fundamentalmente esté de nuestro lado y comparta nuestros valores. No veo cómo evitar este problema difícil.

En realidad soy bastante optimista de que este problema pueda ser resuelto. No tendríamos que escribir una larga lista de todo lo que nos importa, o, peor aún, codificarla en algún lenguaje informático como **C++** o Python, sería una reto imposible. A cambio, crearíamos una **I.A.** que use su inteligencia para aprender lo que valoramos, y su sistema integrado de motivación sería diseñado para defender nuestros valores y realizar acciones que se ajusten a ellos. Así que usaríamos su inteligencia tanto como fuera posible para resolver el problema de la atribución de valores.

Esto puede suceder, y el resultado podría ser muy bueno para la humanidad. Pero no sucede automáticamente. Las condiciones iniciales para la explosión de la inteligencia necesitan ser perfectamente definidas si queremos contar con una detonación controlada. Los valores de la **I.A.** tienen que coincidir con los nuestros no solo en el ámbito familiar, donde podemos comprobar fácilmente cómo se comporta, sino también en todos los nuevos contextos donde la **I.A.** podría encontrarse en un futuro indefinido.

Y también hay algunas cuestiones esotéricas que habría que resolver: los detalles exactos de su teoría de la decisión, cómo manejar la incertidumbre lógica, etc. Así que los problemas técnicos que hay que resolver para hacer este trabajo parecen muy difíciles --no tan difíciles como crear una **I.A.** superinteligente-- pero bastante difíciles. Este es la preocupación: crear una **I.A.** superinteligente es un reto muy difícil y crear una que sea segura implica desafíos adicionales. El riesgo es si alguien encuentra la manera de superar el primer reto sin resolver el otro desafío de garantizar la máxima seguridad.

Así que creo que deberíamos encontrar una solución al problema del control por adelantado, de modo que esté disponible para cuando sea necesario. Puede ser que no podamos resolver por completo el problema del control de antemano porque tal vez, algunos elementos solo pueden ser desarrollados después de reunir los detalles técnicos de la **I.A.** en cuestión. Pero cuanto antes solucionemos el problema del control, mayores serán las probabilidades de que la transición a la era de las máquinas inteligentes vaya bien.

Esto me parece algo digno de hacer y puedo imaginar que si las cosas salen bien, la gente en un millón de años discutirá nuestro siglo y dirá que posiblemente lo único que hicimos bien y mereció la pena fue superar con éxito este reto.

Gracias.